Dispensa di Statistica

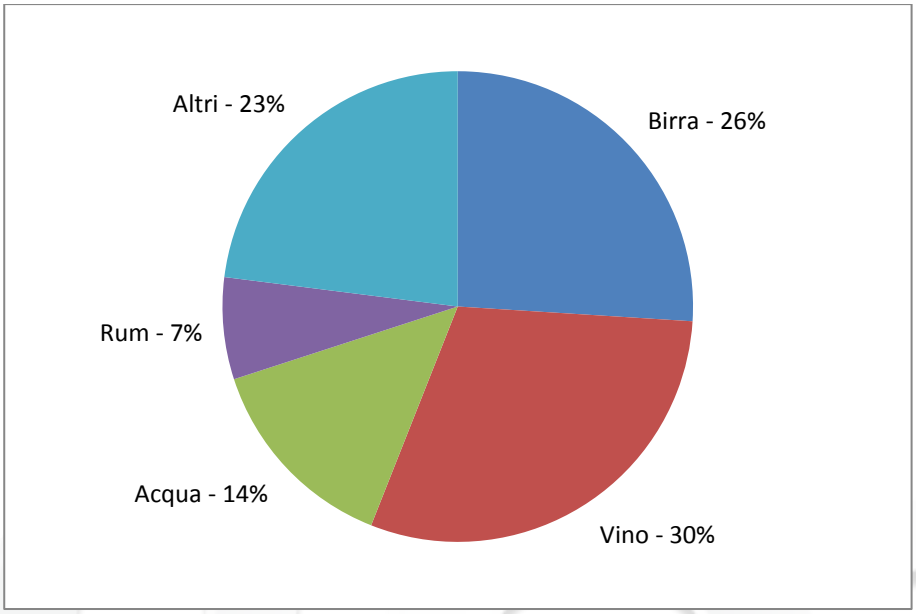
1° parziale 2012/2013

Questa dispensa ha lo scopo di semplificare l'apprendimento dei concetti e delle formule principali. Per approfondimenti contattare l'autore all'indirizzo jackwhile@yahoo.it

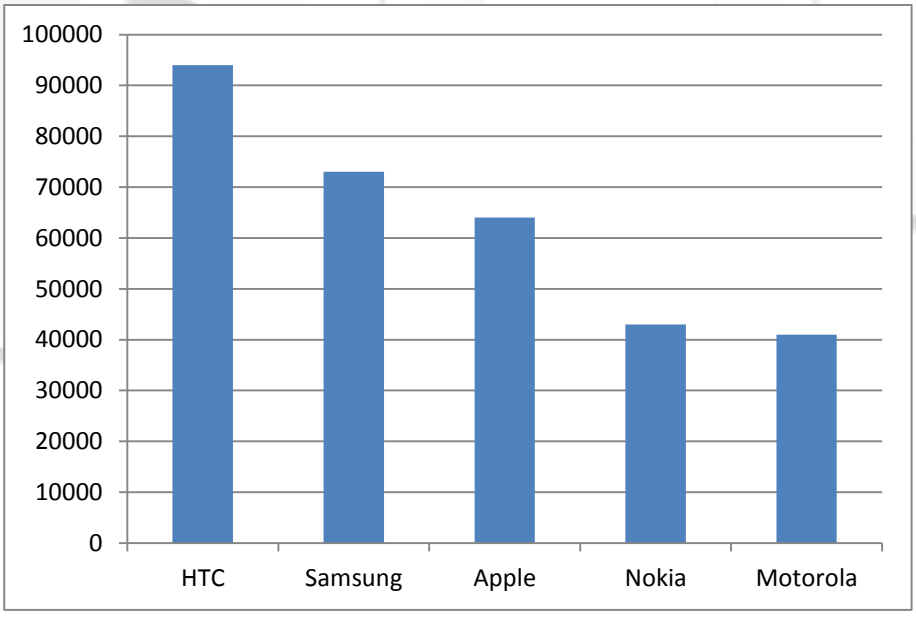
Diagrammi

A seconda del tipo di dato con il quale abbiamo a che fare possiamo fornire una rappresentazione grafica.

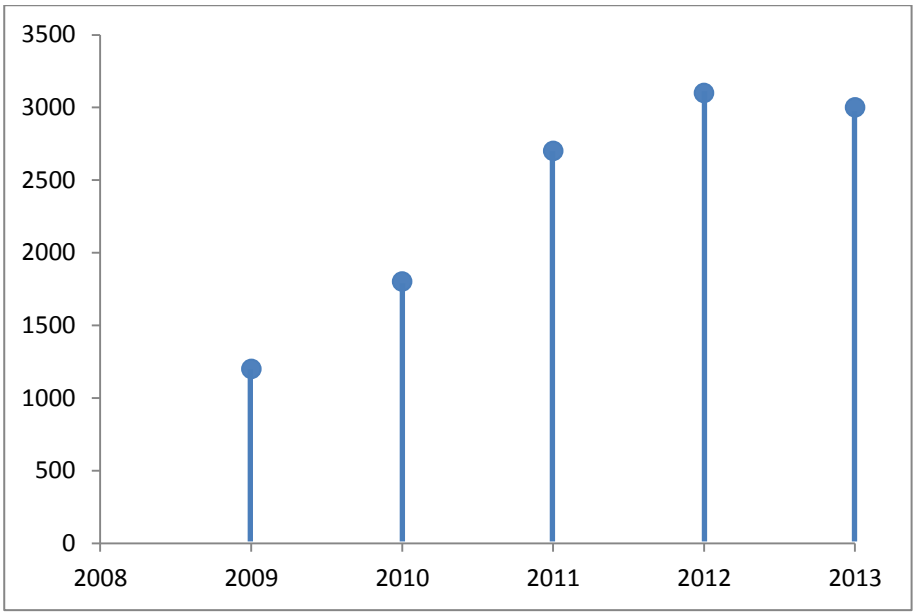
Se la variabile è categorica avremo o il diagramma a torta:



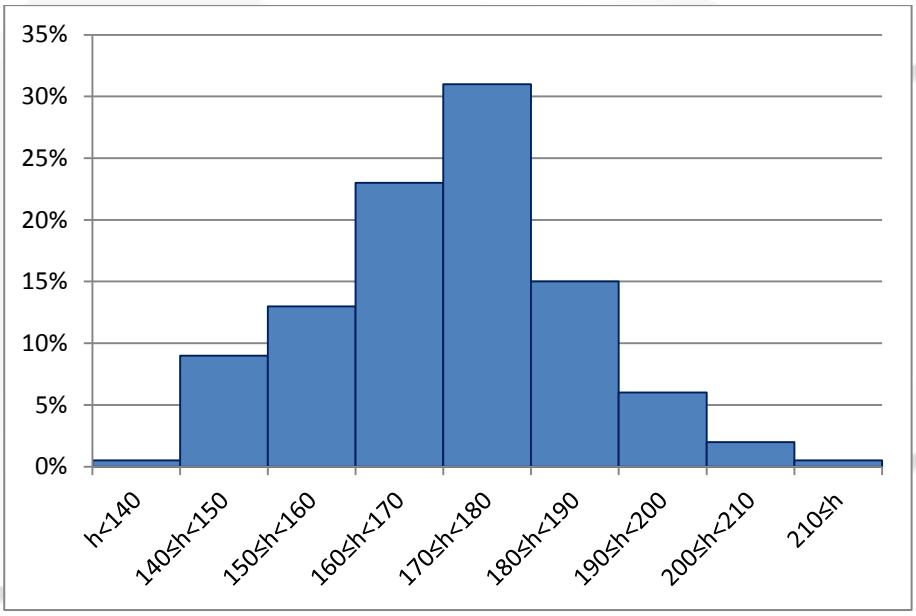
o quello a barre:



Poi ci sono gli altri casi ovvero per caratteri quantitativi discreti:



E per quantitativi continui si usano gli istogrammi



Le frequenze che si possono utilizzare sono sia le assolute che le discrete.

Non è questa la discriminante per la scelta del tipo di grafico.

Indici di posizione

Nella statistica descrittiva i dati sono noti.

Non dobbiamo fare previsioni.

Dobbiamo raccogliere i dati a seconda della loro tipologia, tramite diagrammi appena visti e indici.

Gli indici di posizione centrale sono:

- Media
- Moda
- Mediana

Media

La media, che si calcola solo per variabili quantitative:

$$\mu = M(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

per dati non aggregati (ovvero un semplice elenco di x)

Se abbiamo dati aggregati vuol dire che lo stesso valore non si presenta solitario, bensì se è ripetuto mettiamo, o abbiamo, la frequenza con la quale si ripete.

Se abbiamo le frequenze relative:

$$\mu = M(x) = \sum_{i=1}^{n} x_i f_{x_i} = x_1 f_{x_1} + x_2 f_{x_2} + \cdots + x_n f_{x_n}$$

Mentre se abbiamo le frequenze assolute:

$$\mu = M(x) = \frac{1}{n}\sum_{i=1}^{n} x_i n_i = \frac{1}{n}(x_1 n_1 + x_2 n_2 + \cdots + x_n n_n)$$

Se invece è suddivisa in classi, con classi assolute:

$$\mu = M(x) = \frac{1}{N}\sum_{i=1}^{k} m_i n_i$$

mentre con classi relative:

$$\mu = M(x) = \sum_{i=1}^{k} m_i f_i$$

dove m_i è il punto centrale.

Moda

La Moda è il carattere più frequente

Va bene guardare sia la frequenza assoluta sia quella relativa.

Il valore associato alla frequenza più alta si chiama moda.

Se c'è un solo valore la variabile si dice unimodale.

Se le frequenze più elevate sono uguali, la variabile si dice bimodale quando 2, trimodale quando 3, e così via.

Mediana

La Mediana è il valore prima e dopo il quale sta il 50% dei dati.

Come si trova?

Si ordinano i dati in ordine crescente e si prende il valore che sta nella posizione

$$\frac{n+1}{2}$$

Dove n è pari al numero totale dei dati.

Se n è dispari si prende il valore centrale.

Se n è pari si calcola la media dei 2 valori centrali

La mediana è detta anche Q_2 o secondo quartile.

Il primo quartile si trova con

$$\frac{n+1}{4}$$

Il terzo quartile si trova con

$$\frac{3}{4}(n+1)$$

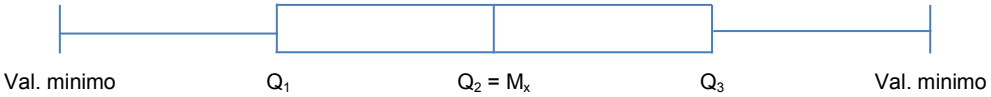
Le 5 misure di sintesi sono:

Valore minimo, Valore massimo, Q_1, Q_2, Q_3

Indici che si ricavano con i quartili sono:

- la media interquartile = $\frac{Q_1+Q_3}{2}$
- il range interquartile = $Q_3 - Q_1$

Una rappresentazione molto usata è il box plot



Se, come in questo caso, $Q3 - Q2 = Q2 - Q1$, allora la distribuzione si dice simmetrica.

Se $Q3 - Q2 > Q2 - Q1$, allora la distribuzione si dice obliqua a destra:

Se infine Q_3-Q_2<Q_2-Q_1, allora la distribuzione si dice obliqua a sinistra:

Indici di dispersione

Varianza

La varianza della popolazione per dati non aggregati è:

$$var(x) = \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

La varianza della popolazione per dati aggregati è:

$$var(x) = \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)n_i$$

Per un campione abbiamo la varianza campionaria; per dati non aggregati:

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Mentre per dati aggregati:

$$S_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\overline{x}^2\right)$$

Scarto Quadratico Medio (SQM)

Lo scarto quadratico medio serve a dirmi di quanto mi discosto dalla media. È detto anche *deviazione standard*. È pari alla radice quadrata della varianza.

Dati i casi visti prima, lo scarto quadratico medio per dati raggruppati è:

$$\sigma_x = \sqrt{\sigma_x^2}$$

Mentre quello campionario è:

$$S_x = \sqrt{S_x^2}$$

Il campo di variazione è banalmente:

$$V_{max} - V_{min}$$

Il coefficiente di variazione è:

$$\mathrm{CV} = \frac{\sigma_x}{|\mu_x|}$$

per una popolazione, ed è:

$$\mathrm{CV} = \frac{S_x}{|\bar{x}|}$$

per un campione.

La disuguaglianza di Chebycheff

Fornisce un limite inferiore di probabilità

$$Fr(\mu - k\sigma < x < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

Data una certa popolazione, k è una costante > 1

Quando la distribuzione è simmetrica, valgono le regole empiriche

Per $k = 1$ $\quad Fr(\mu - \sigma < x < \mu + \sigma) \cong 68\%$

Per $k = 2$ $\quad Fr(\mu - 2\sigma < x < \mu + 2\sigma) \cong 95\%$

Per $k = 3$ $\quad Fr(\mu - 3\sigma < x < \mu + 3\sigma) \cong 99{,}73\%$

Quando abbiamo due variabili in gioco entrano in campo altre grandezze

Covarianza di una popolazione

$$Cov(xy) = \sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

dove:

x_i è il generico valore osservato tra le N osservazioni della variabile x

y_i è il generico valore osservato tra le N osservazioni della variabile y

N è il numero totale di osservazioni

Covarianza campionaria

$$Cov(xy) = S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

dove:

per x_i e y_i vale il discorso visto sopra

$\overline{x}$ e $\overline{y}$ sono le medie campionarie corrispondenti

n è la dimensione del campione

Un indice importantissimo per stabilire l'entità di una relazione lineare fra 2 grandezze è il

Coefficiente di Correlazione Lineare

Esso è dato dal rapporto tra Covarianza e il prodotto degli Scarti Quadratici Medi di x e y rispettivamente:

$$\rho = \frac{Cov(xy)}{\sigma_x \sigma_y}$$

Nel caso di un campione si ha il Coefficiente di Correlazione Lineare Campionario

$$r = \frac{S_{xy}}{S_x S_y}$$

dove

S_{xy} è la covarianza campionaria

S_x ed S_y sono gli scarti quadratici medi campionari

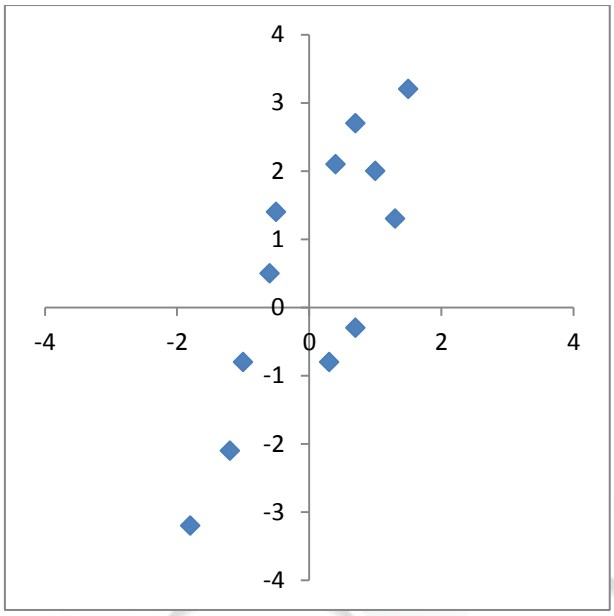
Una regola pratica per riscontrare una relazione lineare è la verifica della seguente:

$$|r| > \frac{2}{\sqrt{n}}$$

Vediamo ora dei grafici che rappresentino o meno la presenza di una relazione lineare

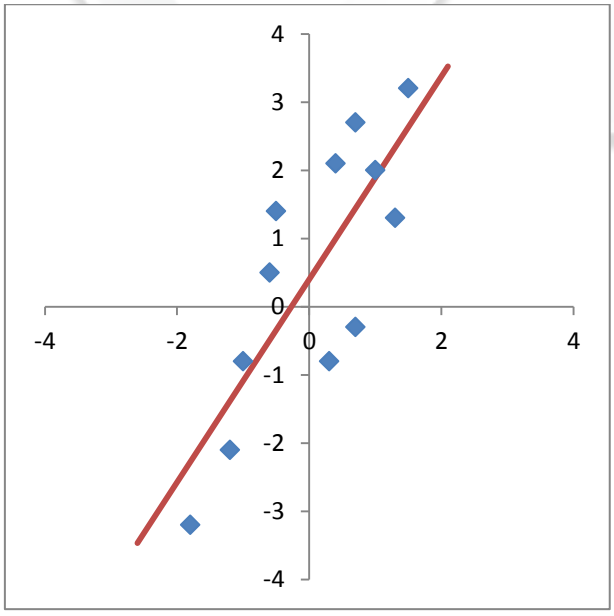
Si ricordi innanzitutto che $-1 \leq \rho \leq 1$, dove $\rho = -1$ indica perfetta correlazione lineare negativa, mentre $\rho = 1$ indica perfetta correlazione lineare positiva.

La densità dei punti ci fornirà indicazioni in merito:

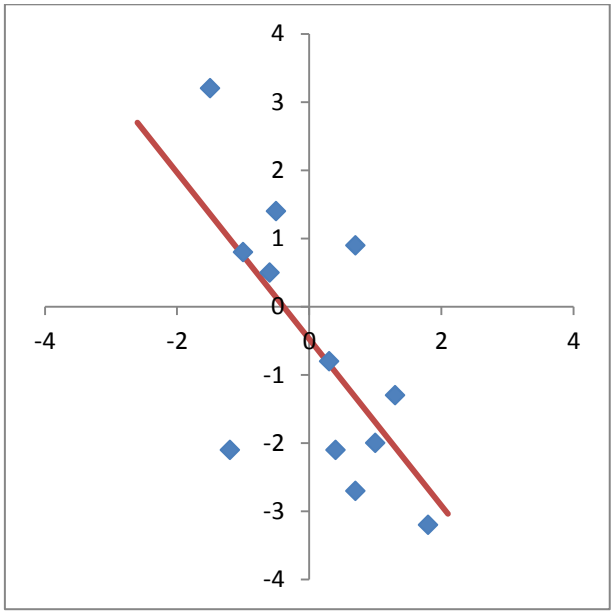


In questo caso i punti si addensano in maniera significativa, con $\rho = 0{,}8$.

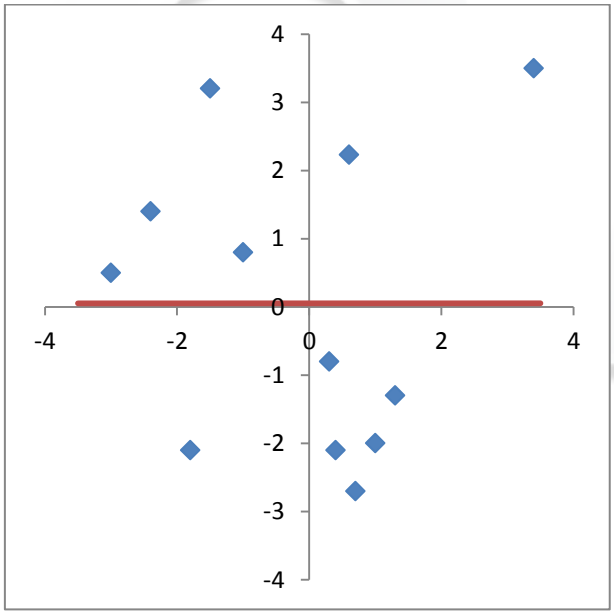
Possiamo quindi disegnare una retta che abbia inclinazione positiva.



In caso invece di correlazione negativa ($\rho = -0{,}66$):



Se $\rho = 0$ non vedo né una retta con pendenza positiva né una con pendenza negativa.



Modello di regressione lineare

È un modello che ipotizza una relazione lineare tra 2 grandezze x e y

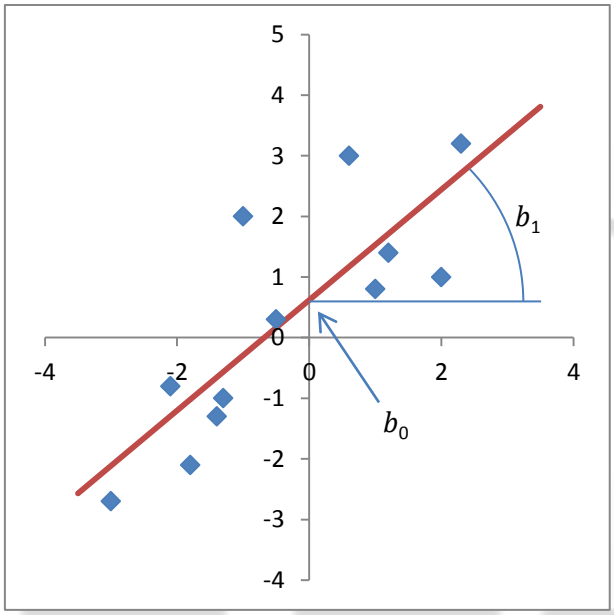
$$y = \beta_0 + \beta_1 x$$

dove y è la variabile dipendente e x è la variabile indipendente.

Assunti b_0 stimatore corretto per β_0 e b_1 stimatore corretto per β_1, si può scrivere che il valore stimato di y è:

$$\hat{y} = b_0 + b_1 x$$

dove b_0 è l'intercetta e b_1 è la pendenza:



Vediamo come si calcolano:

$$b_1 = \frac{Cov(xy)}{S_x^2} = r\frac{S_x}{S_y}$$

e

$$b_0 = \bar{y} - b_1\bar{x}$$

La bontà o meno della retta si vede da

$$0 \leq R^2 \leq 1$$

Se $R^2 = 0$ il modello non è significativo, non è utilizzabile di fatto.

Se $R^2 = 1$ le stime che abbiamo di y sono perfettamente descritte dal modello.

Valori di R^2 superiori a 0,85 indicano un buon modello.

Valori di R^2 inferiori a 0,2 indicano che il modello dà stime poco significative.

$R^2 = \rho_{xy}^2$ nelle popolazioni

$R^2 = r_{xy}^2$ nei campioni

Essendo un potenza pari, R^2 non può assumere valori negativi.

Esso è il quadrato del coefficiente di correlazione lineare.

Stimatori

Uno stimatore si dice corretto se il valore atteso di tale stimatore è uguale al parametro da stimare

Esempio: T_n è lo stimatore del parametro θ da stimare

$$E(T_n) = \theta$$

Esso è asintoticamente corretto se:

$$\lim_n E(T_n) = \theta$$

Uno stimatore corretto è detto anche *non distorto*, ovvero a distorsione nulla.

La distorsione di un generico stimatore T_n si calcola:

$$D_{T_n} = E(T_n) - \theta$$

Ovviamente, se è corretto, $E(T_n) = \theta$, cioè $D_{T_n} = 0$

Uno stimatore corretto per la media di una popolazione normale è la media campionaria $\bar{x}_n$ ha le seguenti caratteristiche:

$$E(\bar{x}_n) = \mu$$

$$\sigma^2_{\bar{x}_n} = \frac{\sigma^2}{n}$$

$$\sigma_{\bar{x}_n} = \frac{\sigma}{\sqrt{n}}$$

Teorema centrale del limite

Dato un set di variabili aleatorie $(x_1, x_2, \dots, x_n)$ i.i.d. (indipendenti e identicamente distribuite), con una media μ e varianza σ^2, il teorema centrale del limite afferma che per n abbastanza grande, ogni distribuzione può essere considerata come una Normale.

Data una proporzione campionaria

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

si ha una distribuzione che può essere considerata una normale standard con buona approssimazione, se

$$n \cdot p \cdot (1 - p) > 9$$

Stimatore efficiente

Ha un significato non tanto assoluto quanto relativo.

In statistica uno stimatore migliore si dice più efficiente.

Ad esempio, T_1 è più efficiente di T_2 se il suo errore quadratico medio è inferiore:

$$EQM_{T_1} < EQM_{T_2}$$

Come si calcola:

$$EQM_{T_1} = Var(T_1) + D_{T_1}^2$$

$$EQM_{T_2} = Var(T_2) + D_{T_2}^2$$

Se uno stimatore è efficiente, ovvero non distorto, è chiaro che:

$$EQM_{T_x} \equiv Var(T_x)$$

poiché $D_{T_x} = 0$

Se vi sono più stimatori non distorti, è più efficiente quello con la varianza più piccola.

θ_1 è più efficiente di θ_2 se:

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$

Un altro parametro da tenere in considerazione è l'Efficienza Negativa:

$$EN = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$$

Intervallo di confidenza

per la media di una distribuzione normale con varianza della popolazione nota:

$$IC_{1-\alpha}(\mu) = \left[\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

$$Margine\ di\ Errore = ME = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

$$ampiezza\ dell'intervallo = w = 2 \cdot ME$$

Per trovare il valore di $z_{\frac{\alpha}{2}}$ si usano le tabelle della normale standardizzata.

$1 - \alpha$ è il livello di confidenza, mentre α è il livello di significatività

L'intervallo di confidenza per la media di una popolazione distribuita normalmente con varianza NON NOTA

$$IC_{1-\alpha}(\mu) = \left[\bar{x} - t_{\frac{\alpha}{2}}^{n-1}\frac{S}{\sqrt{n}}; \bar{x} + t_{\frac{\alpha}{2}}^{n-1}\frac{S}{\sqrt{n}}\right]$$

dove $n - 1$ sono i gradi di libertà, S è la deviazione standard campionaria

Per trovare il valore di $t_{\frac{\alpha}{2}}^{n-1}$ si usano le tabelle della t-student

Il margine d'errore:

$$ME = t_{\frac{\alpha}{2}}^{n-1}\frac{S}{\sqrt{n}}$$

Intervalli di confidenza per la proporzione (grandi campioni)

$$IC_{1-\alpha}(p) = \left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

che è l'intervallo di confidenza per una Bernoulliana.

Il relativo margine d'errore:

$$ME = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Bernoulliana

Ampiezza dell'intervallo $L = 2 \cdot ME$

La numerosità del campione per una normale

$$n = \left(\frac{z_{\frac{\propto}{2}}\sigma}{ME}\right)^2$$

La numerosità per una Bernoulliana

$$n = 0{,}25\left(\frac{z_{\frac{\propto}{2}}}{ME}\right)^2$$